Use of Similarity Measures in Elastic K-Means Clustering

Yogita K. Patil
P.G. Student
SSBT's COETJalgaon M.S.
yogitapatil1007@gmail.com

Sandip S. Patil
Associate Professor
SSBT's COET Jalgaon M.S.
sspatiljalgaon@gmail.com

Abstract

Clustering is used to assign given data to a proper cluster, there are various types of clustering techniques. Standard K-means technique assigns a data points in a single cluster, but In real world there is an ambiguous noisy data and clustering should assign such data to proper cluster. Standard K-means technique cannot handle the ambiguity in the data. Elastic K-means can handle ambiguous data as it uses vector of attributes, but ambiguous data has both vector and similarity. In this paper we are proposing a feasibility of extending the performance of elastic k-means using similarity measures, so that it can handle the real world ambiguous data.

Keywords: *Elastic K-means clustering, Word sense disambiguation, Ambiguity.*

1. Introduction

Clustering is common technic use for describe data into groups or cluster. This grouping is based on features or attributes of the data set. From these attributes and features clustering is find the similarities between them and group them into one same cluster. All clusters are different than each other. Clustering has two types hard clustering and soft clustering. Hard clustering group the data set into 1 and 0 and soft clustering have flexible grouping as compare to hard clustering technic. Fuzzy clustering is soft clustering technic.

Fuzzy clustering partition the cluster and it provide automatic detection of cluster boundaries and many times boundaries get overlap is depends on situation. Membership degree also improves into fuzzy clustering. Because data in data set is belongs to more than one cluster [1][3][5]. Many important features of fuzzy clustering like membership value provided by fuzzy clustering which are useful for sense also it is flexible etc.

1.1. Elastic K-means Clustering

Fuzzy K-means deal with the feature vectors of data. In AI applications, alongside feature vectors, in data points various pairwise relations which are expressed as graph data. Existing clustering uses graph data, such as Normalized Cut based methods, Min and Max Cut based methods and Ratio Cut based methods. Clustering methods uses nonlinear method IsoMAP and linear embedding method PC and Local linear Embedding (LLE) to embed graph nodes in low-dimensional space. Soft capability posterior probability is used by EKM clustering, in which each data point belongs to multiple clusters. Information is used to get the final results. Elastic K-means is little bit different than fuzzy K-means algorithm. Elastic k-means clustering make clusters as per the similarity of the data set, but all clusters are different than each other. Elastic K-means uses posterior probability. Posterior probability it not used into k-means algorithm. It is a uniqueness of Elastic k-means algorithm. Important observation in elastic k-means clustering is that each data point is assigned to the clusters scattered into several possible classes according to its posterior probabilities. To show improvement of clustering using elastic k-means clustering we used some datasets.

2. Literature Survey

Ginter, F., Boberg, J. and Järvinen, J. [2] proposed new techniques for disambiguation in natural language and their application. It utilized supervised SVM machine learning method based on the weighted bag-of- words. Their approach increased in accuracy from 79% to 82%.Thanh Le et al., in [4], introduce Fuzzy C-means algorithm. . Fuzzy clustering is soft clustering technic. It allows allows data objects to belong to multiple clusters based on the degree of membership. Data item are not equally distributed intoclusters and cluster size also different. Hence clustering get opportunity to deal with the data that belong to more than one cluster at the same time.

Zhang, D. Q., Chen, S. C. [6] proposed clustering incomplete data using kernel based fuzzy c-means algorithm. The KFC method can map the initial data into a high-dimensional feature space. It use the kernels based mercer theorem. Adopting the kernel method can decrease computing time. In this algorithm input data corresponding with the high-dimensional feature space data, based on Mercer theorem. Tapas, D.M.Mount, N. Netanyahu, C. Piatko, R. Silverman, and A.Y.Wu [7] proposed K-means clustering.K-means method used to partition a data set into clusters utomatically. Data in onecluster is same as data in same cluster but thee data members in same cluster are totally different than other cluster data members. K-means offered uncertain boundaries to construct the cluster. Clustering relaxes the requirement by providing gradual membership which is suited for the real world problems.

E. G. Mansoori [8] proposed the fuzzy rules base clustering use for input dataset to form clusters by using the inference. It design cluster automatically based on labeled data. Classification performance and performance are important factors in it. [6]. AihuaZheng, Bo Jiang, Yan Li, Xuehan Zhang and Chris Ding [8] introduce Elastic k-means algorithm. Elastic k-means clustering make clusters as per the similarity of the data set, but all clusters are different than each other. Elastic K-means uses posterior probability. Posterior probability it not used into k-means algorithm. It is a uniqueness of Elastic k-means algorithm. Important observation in elastic k-means clustering is that each data point is assigned to the clusters scattered into several possible classes according to its posterior probabilities. S. Ayramo and T. Karkkainen, introduce Partitioning-based clustering method [11]. It divides data items into few pre-specified number of cluster directly. This clustering have many algorithms such as probabilistic Clustering, K-medoid Clustering K-means Clustering and Relocation Algorithms. Richard C. Dubes and Anil K. Jain, [12] proposed Agglomerative hierarchical clustering. It is commonly used for document clustering and is better than K-means and faster. Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey[13] described The Scatter/Gather system uses hierarchical clustering to produce “seeds” for a final K-means phase. that hierarchical clustering with a K-means refinement is essentially a hierarchical-K-mean hybrid that is common to techniques that other people have tried.

3. Proposed work

In proposed work the pre-processing of data is done. Then the pre-processed data is given to modify Elastic K-means algorithm.

3.1Problem Statement

Clustering divides the data set into clusters. These clusters data have larger similarity as compare to other clusters data. Partitioning is based on similarity of data Assigning each data point into exact one cluster as in traditional clustering often causes problems. Fuzzy clustering offered uncertain boundaries to construct the cluster. Fuzzy clustering relaxes the requirement by providing gradual membership which is suited for the real world problems. Elastic K-means is fuzzy clustering algorithm. EKM have more fuzziness as

compare to hard clustering algorithms, but it does not use similarity of the data, it uses only vector (sometimes called attribute) data, so it leads less accuracy and performance. Proposed system is deal with this problem. Proposed EKM uses vector data as well as similarity data. If we use vector data as well as similarity data then we improve the performance of EKM algorithm.

3.2 Objectives

The objectives of the project are:

- To improve the Performance of EKM clustering.
- To improve the accuracy EKM
- To improve the Membership of EKM Clustering.

3.3 Similarity Measure

Typically, the similarity between documents is estimated by a function calculating the distance between the vectors of these documents: two close documents according to this distance are regarded as similar. Several measures of similarity have been proposed. Among these measurements we can quote:

The cosine distance

$$\cos(d_i, d_i) = \frac{\sum_{t_k}[TF \times IDF(t_k, d_j)] * [TF \times idf(t_k, d_j)]}{||d_i||^2 * ||d_j||^2} \quad (1)$$

The Euclidean distance

$$\text{Euclidean}(d_i, d_j) = \sqrt{\sum_1^n (wki - wkj)^2} \quad (2)$$

The Manhattan Distance

$$\text{Manhattan}(d_i, d_j) = \sum_1^n |wki - wkj| \quad (3)$$

3.4 Use of Similarity Measures in Elastic K-Means Clustering

Updated Elastic K-means algorithm solve the problem of existing algorithm.

Algorithm 1: Updated Elastic K-means Algorithm using Similarity Measures and vector Data.
This algorithm uses a similarity measures in EKM clustering
Input: data Y, Z
Steps:

1. Initialize X0.
2. Construct the indicator X: X_{lm}= 1 if xl belongs to cluster m. otherwise, X_{lm}= 0
3. Initially Update the membership function X
4. while not converged

5. $G_{lm} \leftarrow G_{lm} \cdot \sqrt[4]{\frac{(2CG+DGG^TG+GG^TDG)_{lm}}{(2DG+CGG^TG+GG^TCG)_{lm}}}$

Where

A = (|((YT Y)(ZT Z))ik| + ((YT Y)(ZT Z))ik)/2

B = (|((YT Y)(ZT Z))ik| - ((YT Y)(ZT Z))ik)/2

6. Calculate the similarity between the elements of vectors by using cosine Similarity

$$\cos(d_i, d_i) = \frac{\sum_{t_k}[TF \times IDF(t_k,d_j)] * [TF \times idf(t_k,d_j)]}{||d_i||^2 * ||d_j||^2}$$

7. Update the membership function of each data element

G_{ik}

8**.** Assign each data element to maximum membership value

9. End

3.5Architecture

First we give the input to the system as form of data set then next step is tokenization. After that we remove the stop words from data set and then stemming is apply on data set. Then pre-processing file is given to modified EKM algorithm. Following Figure 1 is proposed system architecture.

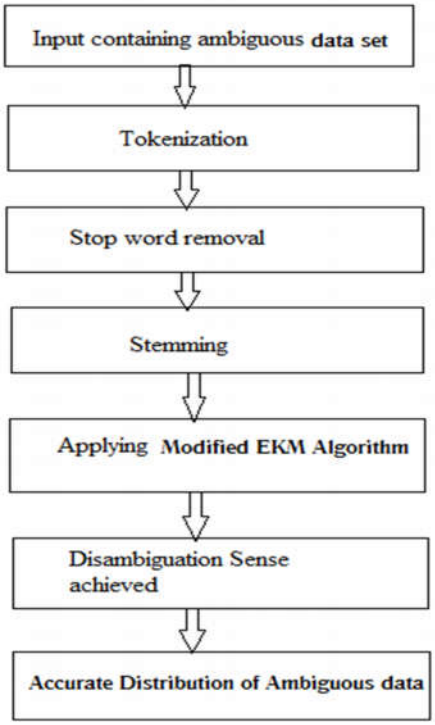

Figure 1: Architecture of Proposed System

In EKM clustering, as an input, it uses vector data as well as similarity data. Clustering is partition the data set as per the features of data like vector data and similarity. The process of elastic K-means is little bit same as Fuzzy K-means algorithm. Clustering

group them as per their context of data. If we compare the result of proposed and existing system then we see deference between them. Hence we modify the Elastic K-means algorithm. From the modification of algorithm performance of EKM clustering algorithm is automatically improved.

Conclusion

Drawbacks of K-means clustering is it assigns data point to one cluster, Elastic K-means resolves it by using flexible membership, but in real data there is a ambiguity. Once we use similarity measures in Elastic K-means, it can accolade such situations. Improved Elastic K-means is able to assigned ambiguous data points into several nearby clusters with the help of vector attributes information and similarity measures.

References

[1] C. T. Baviskar and S. S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", (*2016),* International Conference on Computation of Power energy

[2] Ginter, F.,Boberg, J., Järvinen, J., "New techniques for disambiguation in natural language and their application to biological text", The Journal of Machine Learning Research, vol. 5, (2004), pp. 605-621.

[3] Chitra Liladhar Mahajan, Sandip S. Patil. "Word Sense Disambiguation For Devnagari Language", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Vol.5, Issue 12, pp.447-452, December - 2017

[4] T. Le, T.Altman, and K. Gardiner, "Probability-based imputation method for fuzzy cluster analysis of gene expression microarray data", Ninth International Conference on information technology, Denver, USA, (2012) April.

[5] D. S. Bhole and S. S. Patil, "Detection of paraphrases for Devanagari languages using support vector machine," *2018 International Conference on Communication information and Computing Technology (ICCICT)*, Mumbai, 2018, pp. 1-5.

[6] Zhang, D. Q., Chen, S. C., "Clustering incomplete data using kernel-based fuzzy c-means algorithm", *Neural Processing Letters*, vol. 18 (3),(2003), pp. 155-162.

[7] K. Tapas, D.M.Mount, N. Netanyahu, C. Piatko, R.Silverman, and A.Y.Wu, "An efficient k-means clustering algorithm: analysis and implementation", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 7,Jul 2002, pp. 881-892.

[8] E. G. Mansoori, "Frbc: a fuzzy rule-based clustering algorithm", IEEE Transactions on Fuzzy Systems, vol. 19, no. 5, (2011,)pp. 960-971.

[9] M. S. Yang, "A survey of fuzzy clustering", Mathematical and Computer modeling, vol. 18, no. 11, (1993), pp. 1-16.

[10] Zheng A, Jiang B, Li Y, Zhang X, Ding C, "Elastic K-means using posterior probability", Journal pone, vol. 12, no. 12, (2017).

[11] S. Ayramo and T. Karkkainen, "Introduction to partitioning based clustering methods with a robust example," Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, 2006.

[12] Richard Dubes and Anil K. Jain, "Clustering methodologies in exploratory data analysis", Advances in Computers, vol. 19, (1980), pp.113-228.

[13] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, "*Scatter/Gather: ACluster-based Approach to Browsing Large Document Collections",* Proceeding of the 15th annual international ACM SIGIR conference on research and development in information retrieval, (1992), pp. 318 – 329.